ALEXANDRIA:

Forschung für Web Archive

Stefan Siersdorfer

Forschungszentrum L3S
Hannover

Verlieren wir die Vergangenheit des Web?

EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 أحكام عسكرية من 3 لـ 10 سنوات ضد البلطجة والسرقة بالإكراه وخرق حظر التجوال http://dlvr.it/KCHz8 #tahrir
about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 نشر نموذج بطاقة الاستفتاء على التعديلات الدستورية http://dlvr.it/KCHym #tahrir
about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 عاجل.. حل اتحاد كرة القدم http://dlvr.it/KCHyJ #tahrir
about 1 hour ago via EgyTweets

RT @SoulfunkLA: comes rough, tough like an elephant tusk. Ya head rush, fly like Egyptian musk...
about 1 hour ago via EgyTweets

RT @flavianoflavian: DailyNewsEgypt :Egypt shelled trucks bringing arms from Sudan http://tinyurl.com/4rxo8dx #fb
about 1 hour ago via EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://ow.ly/4e0V4 #Atfeeh #Egypt
about 1 hour ago via EgyTweets

RT @techsynd: Intel Buys Egypt-Based SySDSoft To Boost Its 4G LTE Efforts: http://tinyurl.com/4brmh4n
about 1 hour ago via EgyTweets

Gun running from

Attack on Copts

Spam

ALMASRYALYO
ENGLISH E
MON 14 MAR 2011 LAST UPDATED

News Economy Media Env
21 °C Latest News: Frar

Home » News » Top stories

Armed forces attac

Staff
Mon, 14/03/2011 - 15:23

Like 13
Share
ShareThis

theguardian

News | Sport | Comment | Culture | Business | Money | Life & style

News › Politics › Conservatives

Conservative party deletes archive of speeches from internet

Decade's worth of records is erased, including PM's speech praising internet for making more information available

Randeep Ramesh and Alex Hern
The Guardian, Wednesday 13 November 2013 15.40 GMT

Jump to comments (1284)

A speech in which David Cameron said the internet would help people hold politicians to account was among those deleted. Photograph: Barcroft Media

Verlieren wir die Vergangenheit des Web?

Library of Congress

- In April 2010: Vereinbarung zur Archivierung von Twitter-Daten seit 2006
- January 2013: “It is clear that technology to allow for scholarship access to large data sets is lagging behind technology for creating and distributing such data. The Library is pursuing partnerships to allow some limited access capability in reading rooms.”

Deutsche Nationalbibliothek

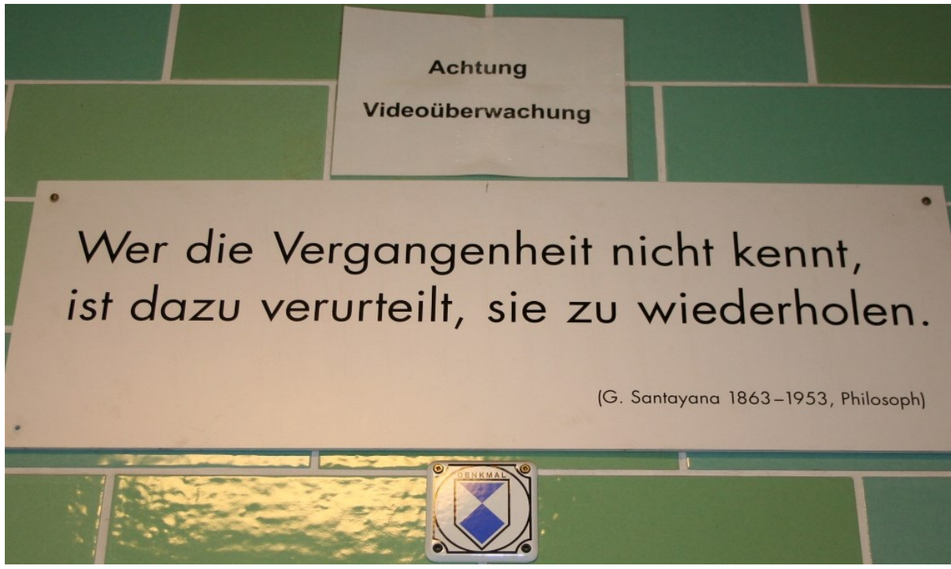
- Gesetzliche Vorgabe seit Juni 2006: sammeln, katalogisieren, archivieren von Web Publikationen

Internet Archive

- Archivierung des Web (10 Petabyte) seit 1996
- Zugang über URLs

Relevante Projekte am L3S

- Web Archivierung: LiWA, ARCOMEM, ForgetIT
- Web Suche: PHAROS, CUBRIK
- Web Analytics: EUMSSI, Qualimaster
- ***ERC Advanced Grant: ALEXANDRIA*** (2014 – 2018, 2.5 Mill. Euro)



Kollaborationen

- Deutsche Nationalbibliothek, British Library, Internet Archive, Rutgers University, et al

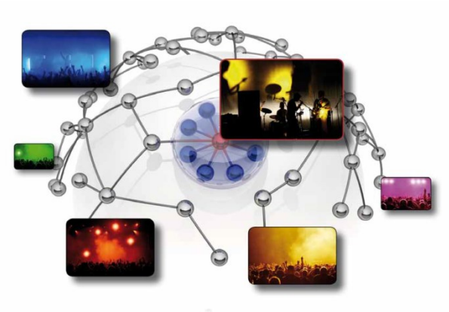
Web Science @ L3S

„Preserving, understanding and shaping the Web”

Informatik- und interdisziplinäre Forschung zu Internet und Web

- Internet: Wie sieht das Netz von morgen aus?
- Information: Wie bekomme ich die Informationen, die ich brauche?
- Community: Wie nutzen Gruppen das Web?
- Society: Welche Anforderungen stellen wir an das Web?

Ausgewählte Projekte



Glocal: Event-basierte Suche

arcomem

Social Web & Web-Archivierung

LivingKnowledge: Meinungsvielfalt im Netz

ForgetIT

ForgetIT: Archivierung, Erinnern, Vergessen

CUbRIK

CUbRIK: Multimedia-Suche für und mit Menschen

QualiMaster

Web-basierte Analyse von finanziellen Entwicklungen

Unser aktuelles Projekt: ERC Advanced Grant ALEXANDRIA

Motivation

- Das Web ist ein Spiegelbild unserer Zeit, vom Alltagsleben bis zur Hochkultur.
- Was bleibt davon in 100 / 1000 Jahren, wenn es niemand bewahrt?

Thema

- Foundations for Temporal Retrieval, Exploration and Analytics in Web Archives

Finanzierung

- 2,5 Mill. Euro für die LUH, 2014 – 2018

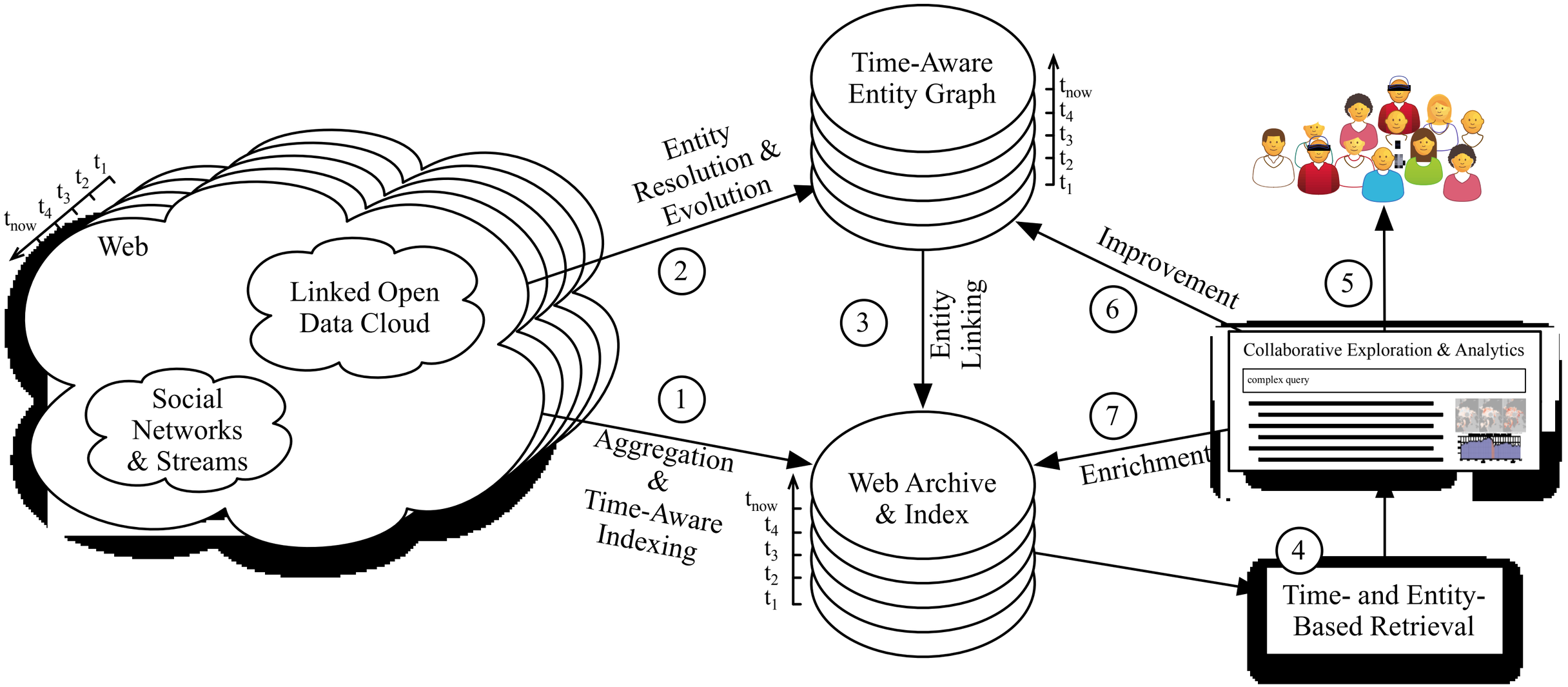
Ziele:

- Bewahrung unseres kulturellen Erbes im World Wide Web
- Entwicklung neuer Modelle und Algorithmen, die es ermöglichen, nicht nur auf die Gegenwart, sondern auch auf die Vergangenheit des Web zuzugreifen.

Kollaborationen u.a. mit

- Deutsche Nationalbibliothek, British Library, Internet Archive, Yahoo! Research Barcelona, Rutgers University

ERC Grant ALEXANDRIA: Temporal Information Retrieval, Exploration and Analytics in Web Archives



ALEXANDRIA Testbeds

Temporal Wikipedia

- Englische, deutsche, italienische Wikipedia mit allen Revisionen
- Links auf Zeitungsarchive (NYTimes, Zeit) und Web-Inhalte
- Entity-Extraktion und Evolution, zeit- and entitäts-orientierte Suche

Webarchive der Forschungsinstitutionen

- Akademische Institutionen in Deutschland und England
- BibSonomy und FreeSearch/DBLP Publikationsdaten
- Entity-Extraktion und Evolution, Exploration und Auswertung

Politik im Web

- Offizielle und inoffizielle Seiten über Politik im Web (Deutschland und England, zusammen mit der Deutschen Nationalbibliothek, British Library, Internet Archive), Stanford US Sammlungen, News, Blogs, Social Media
- Semantische Annotation, Aggregation von Social Media, (journalistische und zeitgeschichtliche) Recherche und Analyse, ebenso wie alle anderen Forschungsfragen

Temporale Suche

query: ncaa

March 14th	March 31th	April 07th
0.0132 · march madness schedule	**0.0100** · oakland raiders	**0.0122** · ncaa women's basketball tournament
0.0117 · ncaa basketball tournament	**0.0090** · ncaaw	**0.0053** · ncaa basketball tournament
0.0068 · nfl draft	**0.0042** · tito francona	**0.0049** · cbs sports line
0.0048 · selection sunday	**0.0031** · ncaa brackets	**0.0033** · ncaaw
0.0037 · oakland raiders	**0.0029** · ncaa division ii	**0.0031** · ncaa final four
0.0032 · 2006 ncaa tournament bracket	**0.0024** · andy goram	**0.0029** · ncaa wrestling
0.0026 · brad hopkins released nfl	**0.0024** · lakers	**0.0028** · march madness bracket
0.0023 · roger clemens	**0.0024** · ncaa women's basketball tournament bracket	**0.0019** · ncaa basketball results
0.0021 · ncaa division ii	**0.0021** · ncaa basketball brackets	**0.0009** · andy goram
0.0014 · college basketball	**0.0021** · nit brackets	**0.0009** · ncaa division ii

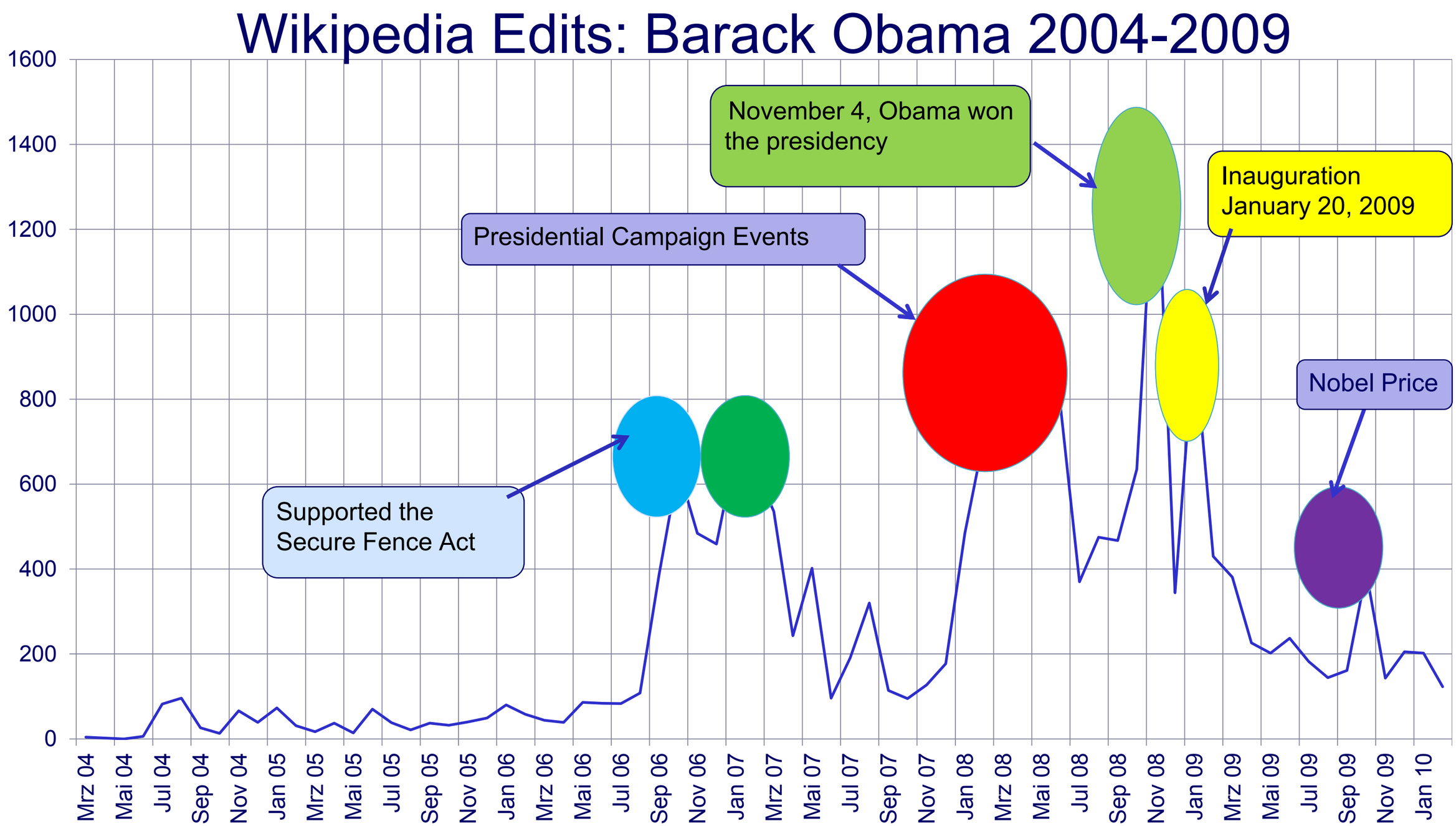
14/03/2006 18/03/2006 01/04/2006

march madness began

ncaa women tournament began

final four began

Extraktion von Ereignissen



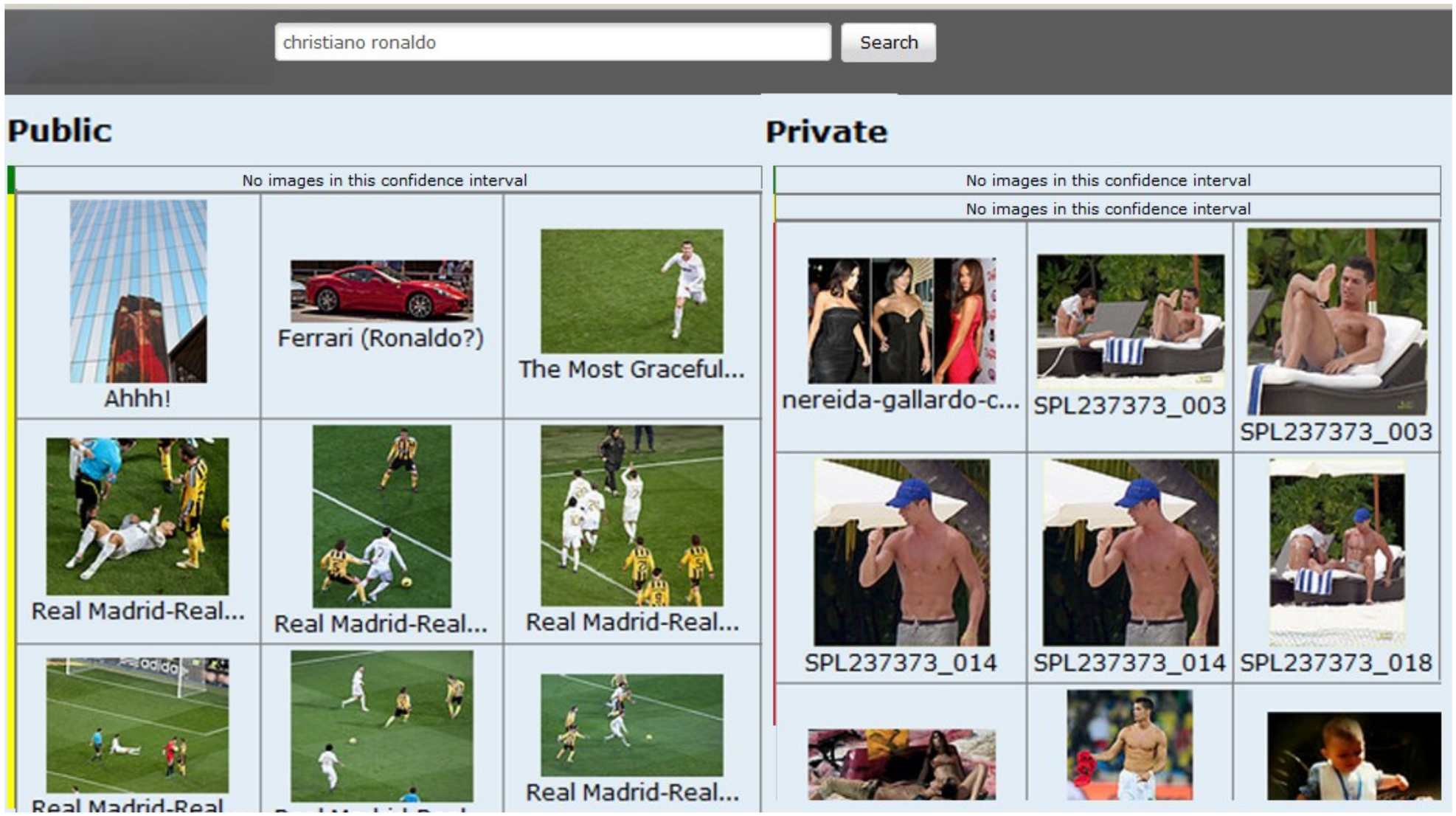
Kommentaranlyse zur Extraktion von Resourcen und Themen

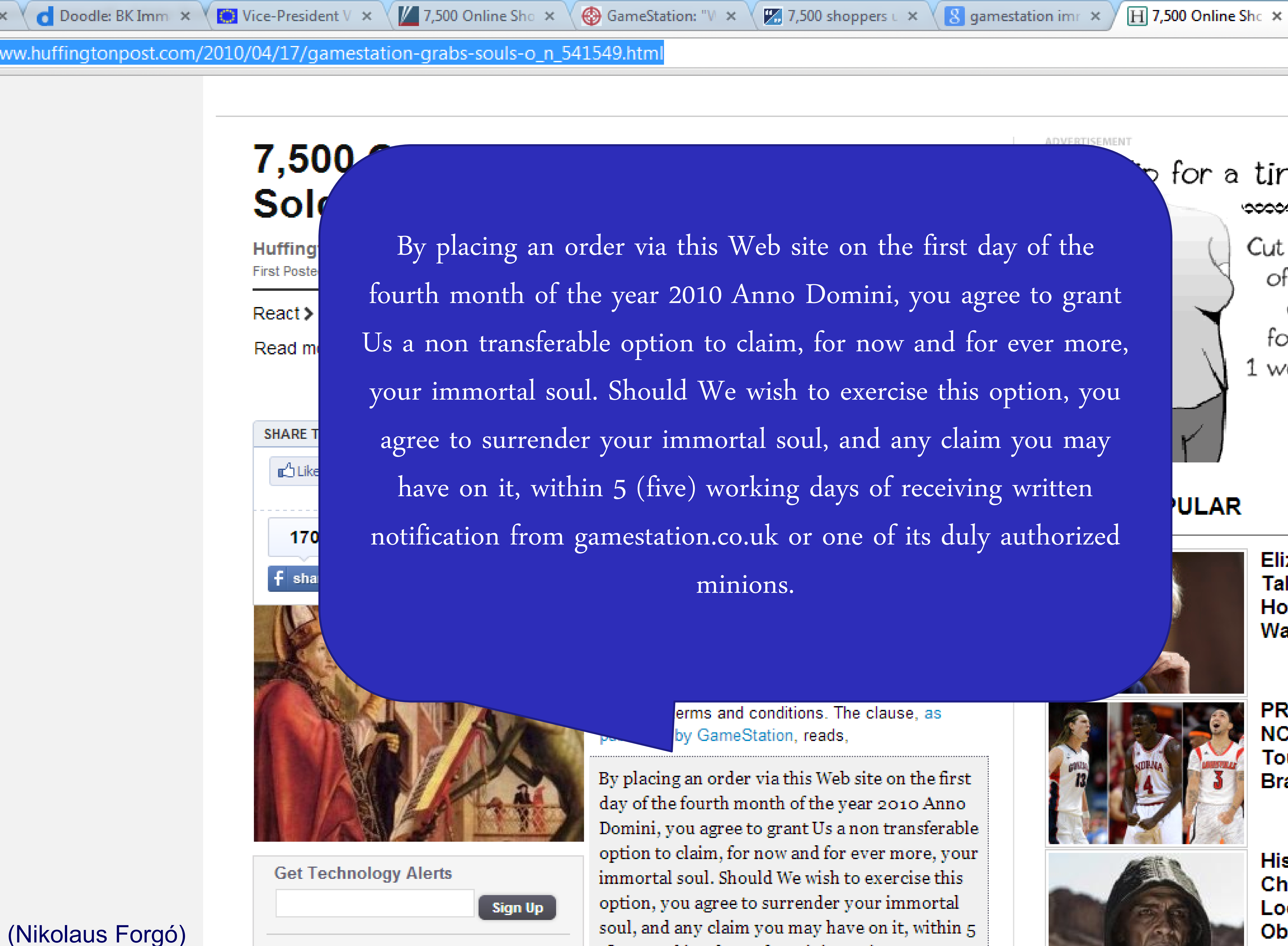
Figure 8: Videos with high (upper row) versus low variance (lower row) of comment ratings

Table 3: Top and Bottom-25 tags according to the variance of comment ratings for the corresponding videos

High comment rating variance				
presidential	nomination	muslim	shakira	islam
campaign	station	itunes	grassroots	nice
xbox	barack	efron	zac	iraq
3g	kiss	obama	deals	celebrities
jew	space	shark	hamas	kiedis
Low comment rating variance				
betting	turns	puckett	tmx	tropical
skybus	peanut	defender	f-18	vlog
butter	chanukah	form	savings	iditarod
lent	daylight	egan	snowboard	havanese
menorah	casserole	1040a	1040ez	booklet

Privacy


christiano ronaldo
Search
Public
No images in this confidence interval
Ahhh!
Ferrari (Ronaldo?)
The Most Graceful...
Real Madrid-Real...
Real Madrid-Real...
Real Madrid-Real...
Real Madrid-Real...
Real Madrid-Real...
Private
No images in this confidence interval
No images in this confidence interval
nereida-gallardo-c...
SPL237373_003
SPL237373_003
SPL237373_014
SPL237373_014
SPL237373_018



(Nikolaus Forgó)

Vielen Dank!

Fragen?